有害表現抽出に対する種単語の影響に関する一考察

畠山 鈴生　桝井 文人　プタシンスキ ミハウ (北見工業大学)　山本 和英 (長岡技術科学大学)

研究背景

ネットいじめ
最近の社会問題 → **ネットパトロール**

- 学校関係者やPTAによるインターネットの監視
- 膨大な書き込み
  →多くの労力と時間が必要
- 特定の人物しか閲覧できないSNS
  →ネットパトロール自体が困難

カテゴリ別関連度最大化手法[1]

フレーズ抽出
係り受け関係を抽出

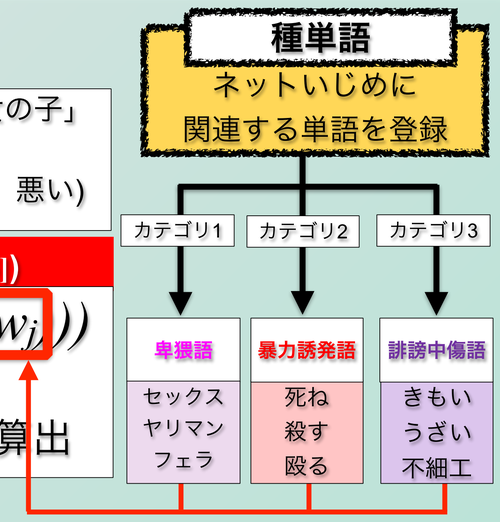
(名詞, 名詞) (名詞, 動詞) (名詞, 形容詞)	例)「性格が悪いかわいい女の子」 ▼ (女の子、かわいい)、(性格、悪い)

WWW検索
種単語とフレーズの関連度を算出

評価モデル(Turney'sの拡張SO-PMI [3])

$$score = max\ (max\ (PMI(p_i, w_j)))$$

フレーズ p_i と種単語 w_j の
*AND*検索により関連度を算出



本研究の位置付け

新田らの手法[1]の性能向上

種単語の規模や組み合わせを考慮 → 種単語による効果の検証

人間解析による精緻化

アンケート画面の一部

対象単語
悪口単語84単語
非悪口単語17単語
Total→101単語

アンケート結果

「有害」	「有害」＋「少し有害」	「無害」＋「少し無害」	「無害」
5	44	30	17

回答者：15名
半数以上(8名以上)を基準として単語を選別

死ね，死ねよ，殺せ，殺す，クズマスゴミ

case1~6との比較で被った単語を選別

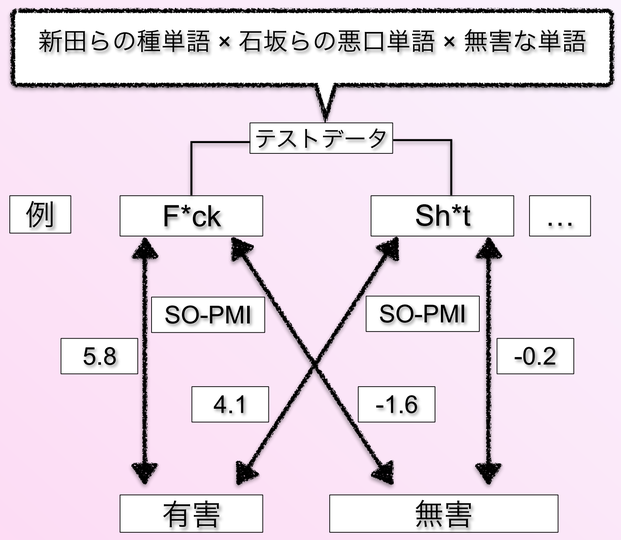
種単語候補

- **case1：7単語**
  17単語の1次フィルタリングで得られた[2]
- **case2：12単語**
  9単語の1次フィルタリングで得られた[2]
- **case3：16単語**
  case1の7単語+元の9単語[1]
- **case4：21単語**
  case2の12単語＋元の9単語[1]
- **case5：5単語 (ベースライン 1)**
  元の単語 [2]
- **case6：9単語 (ベースライン 2)**
  元の単語 [1]
- **case7：5単語**
  「有害」と判断されたアンケート結果
- **case8：18単語**
  アンケート結果とcase1~6との比較

種単語の自動獲得[2]

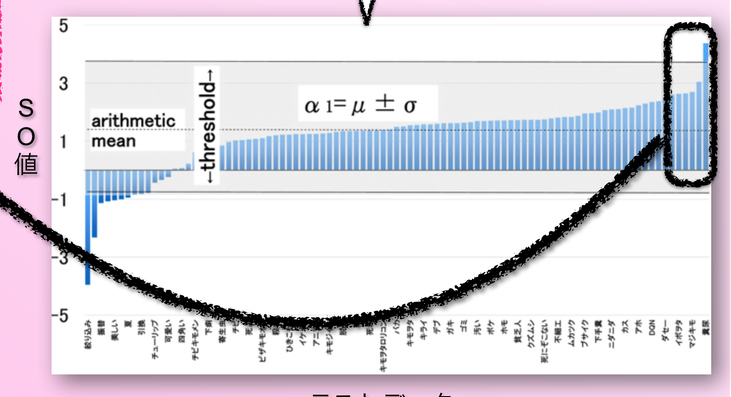
1次フィルタリング

新田らの種単語 × 石坂らの悪口単語 × 無害な単語



2次フィルタリング

石坂らの有害な種単語候補 × フィルタリングした種単語



閾値以上 ＝ 有害

比較実験

表1

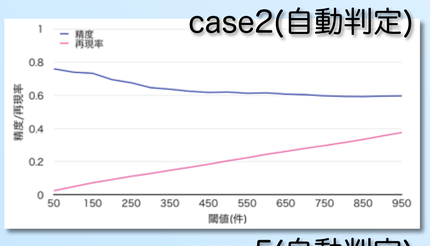
	case1	case2	case3	case4	case5	case6
case5	0.676 vs 0.668 ***	0.676 vs 0.676 ***	0.676 vs 0.671	0.676 vs 0.674	-	0.676 vs 0.674 ***
case6	0.674 vs 0.668 ***	0.674 vs 0.676 *	0.674 vs 0.671 **	0.674 vs 0.674 **	0.674 vs 0.676 ***	-

© Suzuha Hatakeyama 2016　　* p<0.05, ** p<0.01, *** p<0.001

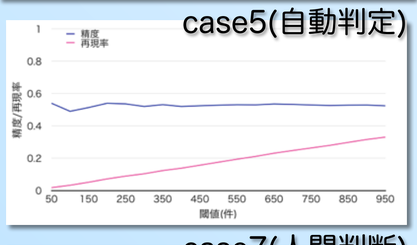
種単語の差による効果を検証
　対象データ ：テストデータ中の1975件
　　(phraseなしを除く)
マクネマー検定によりF値の有意差を確認
帰無仮説：F値に差がない(性能に差がない)

全てのcase(1～4)はcase6よりも性能が高い
case1と2はcase6よりも性能が高い
case5は最も性能が高い

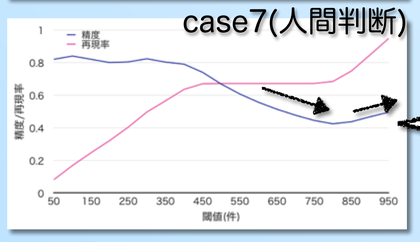
case2(自動判定)



case5(自動判定)



case7(人間判断)



テストデータ
有害：1508件
非有害：1490件
Total：2998件

文脈によって有害にも無害にも変化する単語が存在

	自動検出	vs	人間判断
P	急激に減少		閾値150付近より急減少 800付近で上昇
R	単調増加		閾値450付近まで単調増加, 450～850付近まで一定に 850付近から急上昇

まとめ

種単語の規模は性能に影響するかどうかの検証
→種単語の規模より質が重要

比較実験より
→人手で判断した単語は高い極性値を保持
→精度に影響を及ぼす

今後の課題
文脈により有害にも非有害にもなる文への対応

参考文献

[1] 新田大征，桝井文人，プタシンスキ・ミハウ，木村泰知，ジェプカ・ラファウ，荒木健治：“カテゴリ別関連度最大化手法に基づく学校非公式サイトの有害書込み検出”，第27回人工知能学会全国大会発表論文集，(2013.6).

[2] 石坂達也，山本和英：“Web上の誹謗中傷を表す文の自動検出”，言語処理学会第17回年次大会発表論文集，pp.131-134(2011.3).

[3] Peter D. Turney. 2002. **Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews**. In *Proc. of ACL-2002*, Philadelphia, pp. 417-424, 2002.